Propuesta de investigación: Predicción del valor de una vivienda en Perú según el crédito recibido.

Big Data- Economía- Otoño 2024

Autores: Sophie Schulzen, Vicente Zervino y Sofía Ellenberg

Profesores: Maria Noelia Romero y Victoria Oubiña

Introducción:

Hace ya varias décadas, Perú se encuentra afrontando una gran problemática habitacional. Aunque las viviendas existen y están mayormente construidas, las condiciones en las que habitan muchas personas son precarias. El gobierno no ha dejado de lado este fenómeno y ha implementado diversas políticas públicas con el objetivo de mejorar la calidad de vida de la población peruana. Con una principal concentración poblacional en la capital de Perú, el programa de Barrios Obreros patrocinado por la Dirección de Obras Públicas del Ministerio de Fomento, ha construido cuatro conjuntos ubicados en terrenos de zonas de expansión de la ciudad[1]. Sin embargo, con el gran aumento de la demanda de viviendas en las zonas económicamente activas, el gobierno no logró dar abasto con la construcción directa de viviendas y en respuesta, comenzó a ofrecer planes hipotecarios para facilitar la construcción y adquisición de hogares. Este fenómeno se caracteriza como un problema de exceso de demanda por parte de la sociedad, que exige una mayor cantidad de viviendas, y una insuficiente oferta por parte del Estado, incapaz de abastecer dichas necesidades.

Diversos estudios afirman que un mercado crediticio no sólo es importante para resolver los problemas habitacionales, sino que también es de crucial importancia para fomentar un crecimiento económico nacional. De este modo, Eyzaguirre del Sante, H., y Calderón Seminario argumenta que, "Un mercado hipotecario más desarrollado contribuye al desarrollo del sector construcción y otros sectores vinculados a éste, tales como los de electricidad, agua y alcantarillado, y las industrias productoras de insumos para la construcción", siendo beneficioso para el país en términos macroeconómicos.[2]

Entre otras cosas, un mercado hipotecario habitacional deberá cumplir con dos cuestiones que no siempre son compatibles. Debe poder satisfacer la rentabilidad de los agentes principales del mercado para poder financiar los recursos privados para la

[1]Quispe Romero, J. (2005). *El problema de la vivienda en el Perú, retos y perspectivas*. Revista INVI, 20(53), 20-44. Universidad de Chile. *p 20*. https://www.redalyc.org/pdf/258/25805303.pdf
[2]Eyzaguirre del Sante, H., & Calderón Seminario, C. (2003). *El mercado de crédito hipotecario de Perú*. Instituto Apoyo. Banco Interamericano de Desarrollo. *p 5.* https://cendoc.esan.edu.pe/fulltext/e-documents/MercadoCreditoHipo03.pdf

construcción de las viviendas y las hipotecas, y a su vez, debe resultar factible para aquellos prestatarios, que en particular, se caracterizan por ser el sector con menores ingresos. Siguiendo con esta misma idea, la probabilidad de cumplir con un pago hipotecario depende de una serie de factores, principalmente, de los ingresos de la familia que accede al crédito y de los contactos que dicha familia puede tener. Por esta razón, el mercado tiende a otorgar mayores tasas de interés a aquellas familias de bajo nivel socioeconómico cuya probabilidad de pago es menor, lo que convierte a este mercado en un círculo vicioso para el prestatario. A raíz de esto se pueden deducir tanto ventajas como desventajas. Por un lado, los créditos hipotecarios tienen el gran beneficio de otorgar un acceso inmediato a una vivienda a través de una tasa de interés y cuota mensual estable. Por otro lado, este mercado conlleva el gran riesgo de impago y la posibilidad de perder la vivienda recibida en caso de no cumplir con las condiciones preestablecidas.

Teniendo en cuenta el contexto habitacional de Perú y el mercado crediticio actual, la presente propuesta de trabajo tiene como finalidad analizar diversas cuestiones acerca de esta problemática. En primer lugar, se comentará la existencia de proyectos de investigación relacionados con la temática a estudiar. Luego, se llevará a cabo un análisis acerca de qué tipo de hogares son los que tienen acceso a los créditos hipotecarios a partir de un análisis exploratorio de los datos. Dicho estudio será útil para comprender mejor aquellos factores que se relacionan con el otorgamiento de créditos y con el contexto de la sociedad actual, brindando insumos para lograr diseñar políticas públicas más efectivas. A partir de entonces, evaluaremos un modelo de predicción para el valor de una vivienda en Perú con respecto a ciertas variables relevantes, entre ellas, el tipo de crédito que ha recibido la familia. Más específicamente, utilizaremos los modelos LASSO, Ridge y Naive Elastic Net, para predecir el valor esperado de una vivienda a partir de las características individuales del receptor, el lugar de residencia y el crédito recibido. Finalmente, para concluir el trabajo, desarrollaremos

un modelo de riesgo proporcional que nos permita predecir la probabilidad de que un prestatario logre o no cumplir con el pago del crédito. Este modelo considerará las diferentes tasas de interés recibidas, los distintos montos del crédito, el tipo de prestatario y su historial de crédito.

Literatura Previa:

Basándonos en la propuesta de nuestro artículo resulta crucial analizar las investigaciones existentes que estudian temáticas similares a la nuestra para asegurar que la idea no haya sido previamente explorada. Tras una extensa búsqueda, hemos identificado estudios que abordan cuestiones relacionadas, pero ninguno que cumpla con el objetivo específico de nuestra investigación. Existe una amplia literatura que se centra en la predicción del valor de las viviendas utilizando métodos de machine learning.

Por ejemplo, el artículo "Housing Market Prediction Problem using Different Machine Learning Algorithms: A Case Study" de Jha, S. B., Pandey, V., Bhardwaj, A., y Jha, A. (2020) emplea una variedad de enfoques de modelado, incluyendo LASSO, árboles de decisión y random forest, entre otros, para predecir los precios de las viviendas de Florida entre 2015 y 2019[3]. Este estudio concluye con una comparación del rendimiento de los modelos que ilustra una idea clave: nunca es correcto afirmar la superioridad de un único método para la predicción. En la misma línea, el estudio de Geerts, M., vanden Broucke, S., & De Weerdt, J. (2023) titulado "A Survey of Methods and Input Data Types for House Price Prediction" presenta una extensa revisión de literatura de predicción de precios, examinando 93 artículos publicados entre 1992 hasta 2021[4]. Utiliza análisis de conglomerados para mapear el dominio de la valoración de viviendas, destacando el uso de técnicas y datos innovadores. Estos estudios son relevantes porque abordan la predicción del precio de la vivienda y nos proporcionan una base sólida en términos de metodología y enfoques, los

[3] Jha, S. B., Pandey, V., Bhardwaj, A., & Jha, A. (2020). *Housing Market Prediction Problem using Different Machine Learning Algorithms: A Case Study. arXiv preprint arXiv:2006.10092*

[4] Geerts, M., Vanden Broucke, S., & De Weerdt, J. (2023). A Survey of Methods and Input Data Types for House Price Prediction. *ISPRS International Journal of Geo-Information, 12*(5), 200

cuales adaptaremos y aplicaremos, sumando a nuestra propia contribución, al mercado crediticio peruano.

Por último, el artículo "Policies and Mechanisms of Public Financing for Social Housing in Peru" (2023) tiene como objetivo evaluar el estado actual de oferta y demanda de los programas de vivienda social en Perú[5]. Este estudio analiza la dinámica de los programas de la vivienda social y su impacto en la accesibilidad a la vivienda y en los mecanismos de financiamiento, utilizando datos gubernamentales y encuestas a hogares. Los hallazgos muestran que los programas de vivienda social en el país, como el Fondo MiVivienda, no logran satisfacer la demanda actual, lo que impide que muchos ciudadanos accedan a una menor calidad de vida.

Tras revisar la literatura existente y encontrar estudios sumamente interesantes y relacionados con nuestro tema, nuestra principal contribución será realizar una predicción de precios de viviendas considerando la influencia de créditos específicos otorgados en Perú. Para ello, emplearemos métodos de machine learning algo distintos a los identificados en estos estudios, lo que nos permitirá entre otras cosas, aportar originalidad a nuestra investigación

Datos:

Para llevar a cabo el proyecto, contamos con una extensa base de datos extraída de los datos abiertos del gobierno de Perú[6]. Dicha base incluye información sobre los créditos recibidos por los particulares a nivel individual, con información sobre el tipo de receptor, y el lugar de residencia. Además, en lo que respecta a los créditos, los datos especifican la fecha en la que se otorga, la fecha a la que se pagan las devoluciones, cuantas devoluciones son, la tasa de interés, el monto de la primer devolución, el tipo de institución financiera que lo otorga y que institución financiera lo otorga. Por otro lado, también se cuenta con

[5] Villanueva-Paredes, K. S., & Villanueva-Paredes, G. X. (2023). Policies and Mechanisms of Public Financing for Social Housing in Peru. *Sustainability, 15*(11), 8919.
[6] Base de datos: https://www.datosabiertos.gob.pe/dataset/caracter%C3%ADstica-de-los-cr%C3%A9ditos-otorgados-por-fondo-mivivienda-fmv

información sobre el tipo de prestatario, donde indica si la persona dejó de pagar el crédito, cuando lo hizo y cual es el monto correspondiente al subsidio y el monto del buen pagador. Por último, en cuanto a los datos sobre los individuos que recibieron el crédito tenemos data del documento nacional para identificarlo, la ubicación geográfica de donde vive, el valor de la vivienda, los ingresos de la persona, el trabajo de la persona, una variable dummy que toma valor 1 si hay información de que alguna vez no pagó una deuda en tiempo y forma.

Es importante destacar que al comienzo del desarrollo de la tesis, la página del gobierno de Perú contaba con una base de datos que incluía toda la información mencionada anteriormente. Sin embargo, hoy en día, cuando se realiza el intento de ingresar a la página, se prohíbe el acceso por razones administrativas. Creemos que dicho problema se puede resolver mediante dos posibles formas:

En primer lugar, si lo que está sucediendo es que la página web estaba disponible hasta hace un determinado tiempo y ya no lo está, intuimos que los datos podrían ser escrapeados sin una dificultad mayor.

En segundo lugar, si dicho problema se debe a que la página se encuentra bloqueada para dispositivos que no obtengan una IP de Perú, el problema podría ser solucionado con un VPN.

Metodología:

Análisis preliminar

En primer lugar, para lograr comprender la variabilidad de nuestros datos, sería util realizar un análisis prematuro de las variables explicativas que nos permite leer y entender detalladamente la información con la que contamos. En este contexto, nuestra principal idea para un modelo exploratorio, es realizar un Análisis de Componentes Principales y mirar cuántas dimensiones se necesitan para llegar a explicar una gran proporción de la variabilidad. Esto nos va a mostrar que tan alta sería nuestra multicolinealidad a la hora de

usar luego las variables como predictoras del valor de las viviendas. Sería interesante incluir por ejemplo; índices relacionados con el tipo de hogar, el tipo de crédito y su relación de pago con los créditos (la fecha en que lo reciben, cuanto tardan en pagar, cuando dejan de pagar entre otras). Además, estamos interesados en mirar la dispersión que hay en las variables más importantes. De esta forma, realizaremos un histograma de los ingresos, acompañado de estadísticos descriptivos relevantes, como la media, la varianza, la mediana y algunos cuantiles.

Por otro lado, para interpretar la distribución de los montos de crédito, nos preguntamos cómo es la variabilidad de los créditos y como es su correlación con los ingresos. Con esta última pregunta, sabiendo que los ingresos son una buena proxy del colateral que uno tiene que devolver, intuimos que dicha relación debe ser positiva. Sin embargo, no es obvio con créditos subsidiados en un contexto de análisis de un programa enfocado en el déficit habitacional, y nos parece interesante ver que los datos hablen por sí solos.

Siguiendo con el pensamiento del colateral por el crédito, se hace lugar a otra cuestión. Así como normalmente hay correlación positiva entre colateral y monto prestado, tiene sentido que también haya correlación negativa entre el colateral y la tasa de interés cobrada. Normalmente, el mercado crediticio se maneja con discriminación de precios intentando ser de grado 1, o capaz se podría definir cómo de grado 1 con muchos subgrupos. El problema surge al observar que la discriminación de precios es necesaria generalmente en el mercado crediticio para evitar problemas de *selección adversa* y *moral hazard*[7][8]. Por eso mismo, en el mercado crediticio el precio es la tasa de interés. No es nada revolucionario plantear que las principales características observacionales que determinan la disparidad de tasas de interés son los colaterales y el historial de crédito del prestatario. Sin embargo,

[7] Akerlof, G. A. (1970). "The Market for Lemons: Quality Uncertainty and the Market Mechanism." *Quarterly Journal of Economics, 84*(3), 488-500.
[8] Holmström, B. (1979). "Moral Hazard and Observability." *The Bell Journal of Economics, 10*(1), 74-91.

volviendo al punto anterior, siendo este crédito de carácter subsidiado, cabe la posibilidad de que la relación de la tasa de interés con el ingreso y el historial crediticio no sea la esperada. Es por esto que queremos también ver el coeficiente de correlación entre la tasa de interés y el ingreso; y entre la tasa de interés y una dummy que muestra si el individuo alguna vez no devolvió un préstamo al que se había comprometido antes del Fondo Mivivienda.

Otro análisis previo a nuestra predicción puede ser un mapa de calor mostrando la densidad de créditos otorgados por distrito y por región. El objetivo de este enfoque es tratar de ver si los créditos se otorgaron de forma más bien federal o de manera centralizada. Dicho mapa nos va a ayudar a entender si la variabilidad territorial es alta o no. De no ser alta, podemos intuir que, probablemente la ubicación no va a ser uno de los predictores más importantes de nuestro modelo.

Análisis principal

Modelo predictivo

Para predecir los precios de las viviendas, vamos a usar tres modelos de regularización: un modelo Lasso, un modelo Ridge, y un modelo de Naive Elastic Net. Si bien los tres modelos son muy similares en cuanto a la manera de regularizar, las condiciones de penalidad son muy distintas entre sí. Es decir, funcionan como un modelo de mínimos cuadrados ordinarios con una penalización automática por inclusión de variables irrelevantes dada por un parámetro lambda (o dos en el caso de Naive Elastic Net). Ridge y Lasso tienen una forma funcional distinta en las penalidades. De este modo, Elastic Net combina las penalidades de ambos, por lo que tiene mejor penalización, pero a la vez resulta, generalmente, en mayor varianza.

$$\text{Lasso} \rightarrow R_l(\hat{\beta}) = \sum_{i=1}^{n} (y_i - x_i'\hat{\beta}) + \lambda \sum_{s=2}^{p} |\hat{\beta}_s|$$

$$\text{Ridge} \rightarrow R_l(\hat{\beta}) = \sum_{i=1}^{n} (y_i - x_i\hat{\beta})^2 + \lambda \sum_{s=2}^{p} (\hat{\beta}_s)^2$$

$$\text{Naive Elastic Net} \rightarrow R_{nen}(\beta) = \sum_{i=1}^{n}(y_i - x_i'\beta)^2 + \lambda_1 \sum_{s=1}^{p}|\beta_s| + \lambda_2 \sum_{s=1}^{p}(\beta_s)^2$$

X se entiende como un vector que contiene un 1 en el primer componente y los valores de las variables predictivas en el resto.

En todos nuestros modelos, vamos a tener al valor de una vivienda como variable dependiente y a todas las otras variables numéricas como variables independientes. Vamos a elegir el parámetro lambda (o los parámetros) en cada caso según Cross Validation, eligiendo el que signifique un menor Error Cuadrático Medio en la muestra de testeo del modelo. Luego, compararemos el error cuadrático medio y nos quedaremos solo con el modelo predictivo que menor ECM tenga.

A partir de entonces, es importante darle interpretación al error cuadrático medio para internalizar el poder predictivo del modelo.

Riesgo proporcional

Finalmente, luego de predecir los precios de las viviendas, queda por responder la segunda pregunta de investigación, relacionada con el riesgo proporcional de no pagar un crédito, en base a si uno ya defaulteo algún crédito en el pasado y los valores de los préstamos y tasas de interés. Bajo este enfoque, la idea principal es estimar una ecuación con el método de Cox usando la variable de si se pagó o no el préstamo en tiempo y forma como variable dependiente esta vez. Como variables independientes, se usan las mismas variables que en el modelo de regularización, sacando ahora la variable dependiente y agregando el valor de la casa.

Sin embargo, en este método el análisis de resultados no se presentará en los valores predichos por el modelo o sus estadísticos. En otras palabras, el enfoque ahora estará centrado en tres de los coeficientes del modelo. De esta forma, se observara el signo y la significatividad de los coeficientes correspondientes al historial crediticio, el monto del

préstamo y la tasa de interés. Para concluir se usará ese coeficiente para calcular la función de verosimilitud parcial. Esto nos permitirá, por ejemplo, observar como es el riesgo relativo de que una persona con mal historial crediticio (nótese que la variable es binaria, mal significa tener dummy = 1) haya dejado de pagar en forma el préstamo proporcional al riesgo de una persona con buen historial crediticio (análogo).

La ecuación a estimar es la siguiente:

$h(t/x_i) = h_0(t) \cdot exp(\sum_{j=1}^{p} x_{ij} \cdot \beta_j)$ siendo $h_0(t)$ una función no especificada.

Los parámetros de interés son los 3 $\widehat{\beta_j}$ relacionados a la tasa, el monto y el historial.

Conclusiones y limitaciones:

Gracias a las nuevas tecnologías y metodologías que Machine Learning ha brindado en los últimos años, nuevos análisis de predicción y descriptivos pueden ser generados. En ese sentido, sabiendo que la economía en Perú ha estado en pleno desarrollo en las últimas décadas y junto con los cambios de gobierno, es fundamental analizar cómo las nuevas medidas de créditos hipotecarios han influenciado a la sociedad peruana. De este modo, proponemos un análisis de predicción para el valor de las viviendas ya construidas y los proyectos a construir a partir de diversas variables que correlacionan con dicho precio. Además, a partir de los distintos créditos y tasas de interés, es posible determinar un modelo de predicción de probabilidad de pago de crédito en base a las características del prestatario y del tipo de crédito. Finalmente, esperamos que los modelos establecidos alcancen niveles de Accuracy adecuados para lograr una predicción eficiente. Sin embargo, podrían surgir algunas limitaciones si no logramos obtener completamente la base de datos necesaria, lo que resulta en la ausencia de variables relevantes. En caso de enfrentar esta situación, proponemos realizar encuestas individuales para recopilar información específica sobre las características del prestatario.

Bibliografía:

- Akerlof, G. A. (1970). "The Market for Lemons: Quality Uncertainty and the Market Mechanism." *Quarterly Journal of Economics, 84*(3), 488-500.
- Eyzaguirre del Sante, H., & Calderón Seminario, C. (2003). *El mercado de crédito hipotecario de Perú*. Instituto Apoyo. Banco Interamericano de Desarrollo. *p 5.* https://cendoc.esan.edu.pe/fulltext/e-documents/MercadoCreditoHipo03.pdf
- Geerts, M., Vanden Broucke, S., & De Weerdt, J. (2023). A Survey of Methods and Input Data Types for House Price Prediction. *ISPRS International Journal of Geo-Information, 12*(5), 200. https://doi.org/10.3390/ijgi12050200.
- Holmström, B. (1979). "Moral Hazard and Observability." *The Bell Journal of Economics, 10*(1), 74-91.
- James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An introduction to statistical learning: With applications in Python.* Springer Nature.
- Jha, S. B., Pandey, V., Bhardwaj, A., & Jha, A. (2020). *Housing Market Prediction Problem using Different Machine Learning Algorithms: A Case Study. arXiv preprint arXiv:2006.10092*. https://doi.org/10.48550/arXiv.2006.10092
- ¨Kleemans, M., & Thornton, R. (2023). Fully Promoted: The Distribution and Determinants of Full Professorship in the Economics Profession. *AEA Papers and Proceedings*. 113: 467-472
- Quispe Romero, J. (2005). *El problema de la vivienda en el Perú, retos y perspectivas*. Revista INVI, 20(53), 20-44. Universidad de Chile. *p 20.* https://www.redalyc.org/pdf/258/25805303.pdf
- Sosa Escudero, W., 2021, *Big data*, 7a edición, Siglo XXI Editores, Buenos Aires
- Sosa Escudero, W., 2022, *Borges, big data y yo*, Siglo XXI Editores, Buenos Aires.
- Villanueva-Paredes, K. S., & Villanueva-Paredes, G. X. (2023). Policies and Mechanisms of Public Financing for Social Housing in Peru. *Sustainability, 15*(11), 8919.